

The English Dialects App: The creation of a crowdsourced dialect corpus

Adrian Leemann^{a,*}, Marie-José Kolly^b, David Britain^c

^a Department of Linguistics and English Language, Lancaster University, County South, Lancaster, LA1 4YL, United Kingdom

^b Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zürich, Switzerland

^c Department of English, University of Bern, Länggassstrasse 49, 3012 Bern, Switzerland

ABSTRACT

In this paper, we present the English Dialects App (EDA) and the English Dialects App Corpus (EDAC). EDA is a free iOS and Android app, launched in January 2016 that features a dialect quiz and dialect recordings. For the quiz, users indicate which variants of 26 words they use and the application guesses their local dialect; for the recordings, users can record a short text. The result is EDAC which includes metadata on mobility, ethnicity, age, educational level, and gender. More than 47,000 users from across the UK have indicated dialect variants for these 26 words, and more than 3,500 users have provided audio recordings. Unavoidably, EDAC does not successfully reflect distributions of age, ethnicity, qualification levels, and other parameters found for the UK population given that smartphone-based research reaches a specific stratum of the population. Yet there are also clear benefits to the sampling strategy used – benefits and pitfalls are discussed in this article. Future analyses will provide the most comprehensive understanding of English regional dialect variation since the work of the traditional dialectologists. We showcase two such analyses in this article. EDAC should, we demonstrate, be of interest to researchers in dialectology but also in forensic phonetics.

1. Introduction

The most recent nationwide dialect corpus for England dates back at least half a century and is based on a geographically broad but socially restricted sample of largely non-mobile, older, rural, male speakers (NORMs) (cf. [1]). The age of this corpus and its restricted speaker profile motivated the collection of a contemporary corpus of dialect and acoustic-phonetic data from across England. This new corpus, enabling us to update our knowledge of regional dialect distribution, has, furthermore, a number of important applications beyond dialectology and sociolinguistics, for example in forensic phonetics. In this paper we present the corpus and the collection protocol under which it was created, and demonstrate the potential of its dialectological and forensic applications.

The collection of large multi-locality corpora in dialectology has a venerable tradition that goes back to the 19th century. Traditional studies from that period relied mostly on questionnaires to elicit dialect material in lexis and phonology. Georg Wenker, for example, documented dialects in the late 19th century by distributing some 50,000 questionnaires with 40 test sentences to teachers across Germany. They were asked to transliterate 40 sentences into the local dialect of the

community. Despite its age and methodological advances in the past century, the material collected continues to be used in contemporary dialectological research [2]. Towards the end of the 19th century, Jules Gilliéron sent out fieldworker Edmond Edmont to cycle across France to conduct hundreds of interviews between 1886 and 1900 [3]. This type of fieldwork by Wenker and Edmont typically resulted in linguistic atlases and lay the ground for future work on dialects in Switzerland, Italy, and Spain [4].

In England, meanwhile, the most significant advances in charting the nation's dialects were made by Alexander Ellis (see, especially [5], but also [6]). Like Wenker, Ellis sent out dialect transliteration tasks to people (usually clergy) – principally two short reading passages (one a story, the other a list of sentences) –, and, like Gilliéron, he was fortunate to have a trained phonetician, Thomas Hallam, to travel around the country collecting data and checking the transliterations. In all, data was collected from 1145 places across those parts of the British Isles in which English was the vernacular language in the mid-19th century. No atlas emerged from this endeavor, instead two maps of the islands' main dialect regions (but see [7]). With the exception of Kurath and Lowman [8], based on data collected in 1930, regional dialect documentation only reemerged onto the agenda in England once again after World War

Abbreviations: API, application programming interface; BKA, German Federal Criminal Police Office; BNC, British National Corpus; EDA, English Dialects App; EDAC, English Dialects App Corpus; FRED, Freiburg English Dialect corpus; ICE, International Corpus of English; NORMs, non-mobile, older, rural, male speakers; ONS, Office for National Statistics; SED, Survey of English Dialects

* Corresponding author.

E-mail address: a.leemann@lancaster.ac.uk (A. Leemann).

II. Orton and Dieth [9] and a large team of fieldworkers collected data (on-the-spot phonetic transcriptions of answers to questions, fill-in-the-gap exercises, and so on, in a very long questionnaire) between 1950 and 1961 in 313 localities across England – the Survey of English Dialects (SED). To preserve or at least record ‘the traditional types of vernacular English’ ([9]; p. 14; see also [4]), fieldworkers interviewed mostly NORMs. The impact of the SED on the dialectology of England has been immense. Dialectologists and variationists have drawn upon the data, for example, to provide a historical backdrop for contemporary research (e.g. [10]) as well as to help ascertain the likely dialects spoken by 19th century colonial emigrants (e.g. [11]). Several atlas publications emerged from the SED (e.g. [12]), and because the data collection protocols were so systematically and carefully followed, the data have lent themselves to later computation using dialectometric techniques (e.g. [13]). Despite this, such traditional approaches to dialectology, anchored in the countryside, were criticized for their almost total abandon of the varieties spoken in urban areas. The advent of sociolinguistic approaches to dialectology in the 1960s saw radical changes in data collection methods, especially with respect to the nature of the sample (a wider spectrum of natives from the community were eligible for investigation), the type of data collected – relatively informal conversations within sociolinguistic ‘interviews’ – and the locations of investigation, a shift from rural areas and geographical coverage to the study of single urban locations. For a considerable period at that time, studies of geographical variation were few [1].

Today, the paper and pen techniques of the traditional dialectologists are gradually being replaced by large scale, computer-aided or mobile device-aided surveys. Since the advent and wide availability of computers, large-scale dialectological analysis has been aided by the construction of machine-readable dialect corpora, though it could be argued that none of these provide the systematic coverage of the Survey of English Dialects. Data in the British National Corpus (BNC), which contains 10 million words of spoken English, was categorized into 28 different dialects, but the BNC has limited value for dialectologists since it was not accompanied by audio [14]. A more recent attempt has been made to collect sound recordings from across the UK in the spoken corpus of BNC2014 [15]. The International Corpus of English (ICE) aims to chart different national varieties of English around the world. ICE, however, is not explicitly set up to investigate local or regional dialect *within* the country. The Freiburg English Dialect corpus (FRED [16]) is a 2.5-million-word collection of transcribed oral history recordings (mostly of NORMs) collated from holdings across Great Britain. Further, there is the Helsinki Dialect Corpus [17], another collection of NORMs, mostly from rural East Anglia and the South-west of England. Of these, only FRED comes close to nationwide coverage, however; but its sample was only slightly more recent than that of the SED: “we were looking for material preferably from the 1970s and 1980s, recording older speakers, or from the 1990s, if these recorded very old speakers” ([16]; p. 36).

Computer- or mobile device-based crowdsourced data collection efforts in the UK are in their infancy. MacKenzie et al. [18] studied phonological, lexical, and syntactic variation from around 5000 speakers across England who completed an online questionnaire disseminated by undergraduate students as part of a class module. Vaux's [19] Cambridge Online Survey of World Englishes crowdsources lexical, phonetic, and morphosyntactic variation in World Englishes (including British English) and has been online for more than a decade. The regional data collected is illustrated on maps on the website of the survey, but – to the best of our knowledge – has not been analyzed further or published yet. The maps show particularly high response rates in urban areas of the Southeast (around London) and the Northwest (around Manchester, Liverpool, and Leeds). Social media data, too, are starting to be used to examine regional variation. A first pilot study conducted by Willis [20] used a ten-day corpus of Welsh tweets to examine the pronoun *chdi*. Willis reports a similar distribution to that found by large-scale traditional surveys – however, with much smaller

expenditure of time and money (see also [21,22]). Most recently, Grieve et al. [23] studied two billion words written by one million tweeters. Most tweets were geocoded with longitude and latitude information. They examined 35 lexical items and their 115 regional variants and compared regional distributions to the BBC Voices corpus [24]. They found that the regional variation reported in the Twitter corpus (collected in 2014) largely aligns with the variation found in the BBC voices data (collected in 2004/2005).

But, overall, while research activity on *individual* varieties of British English is undoubtedly healthy, we still know relatively little about contemporary variation at the regional and national level and about how these individual studies mesh together into a supralocal picture of the dialect landscape of the country. In this paper, we present the core functionalities of a free iOS and Android application – English Dialects App (hereafter EDA) – that was developed to generate a contemporary corpus of the English of England. We restricted the app to England because the app's dialect prediction mechanism (i.e. the gamification-approach that motivated users to provide us with their dialect data, see section 2) relied on there being a systematic historical corpus, with consistent coverage of the same variables from the same time period, that could be used as a comparative baseline. This does not exist for the British Isles as a whole – while surveys have been conducted of the linguistic varieties spoken in Wales, Scotland and Ireland, they were conducted at different times, with different methods, and different sets of variables. Using these corpora would distort the dialect prediction mechanism (it would be based on different kinds of data in different places), and systematic comparison would therefore not be possible. Further, in designing the app, we had to avoid overburdening users. Including all, especially Scottish and Irish variants of many of the variables, along with the English ones, would have made the app unwieldy, with too many variants for many of the variables. As will be demonstrated below, however, speakers outside England, too, participated and provided spoken data.

EDA's main functions are, firstly, to locate local and regional dialect characteristics via a quiz which ‘predicts’ users' dialects based on their responses and, secondly, to gather and make available nationwide audio data via users' uploading of self-recorded readings of a short story. Following the motto ‘There's no data like more data’ (cf. [25]), EDA has automatically collected dialect data from more than 47,000 speakers coming from more than 4900 localities across the UK and more than 3500 speakers from the UK have participated in the audio recording functionality over the course of less than one and a half years. Using an app as a dialect guessing tool is not new: it caught the public's interest in German-speaking Switzerland in early 2013 [26] and in the United States in late 2013, when the New York Times published a web-app – the ‘Dialect quiz’ [27] to predict a user's American English dialect. The US quiz consists of 25 questions such as, ‘What do you call it when rain falls while the sun is still shining?’. The user provides their answer and proceeds to the next question. In the end, dialect location predictions are displayed. Posted on the Times's website within the last 10 days of 2013, this quiz became the year's most popular piece of content [28].

The first part of our paper is devoted to presenting the core functionalities of the English Dialects App: the dialect guessing quiz and the dialect recordings. The second part presents descriptive statistics of the outcome of this automated data collection, the English Dialects App Corpus (EDAC). We then discuss the use of EDAC for research and showcase some early results on language change and population statistics of acoustic parameters for the UK. It will become clear that the development of such a database is key to the discovery and quantification of linguistic phenomena on a hitherto unprecedented scale. We end the article with a discussion of the challenges and benefits of EDAC.

2. English Dialects App (EDA)

EDA's core functionalities were driven by the incorporation of

Table 1
Variables chosen for the dialect quiz, prompts, example variants, variant count, and variable type.

| Variable | Prompt | Example variants | N | Type |
|---|---------------|------------------------|----|---------------|
| Lexical variation in <i>autumn</i> | autumn | autumn, fall | 3 | Lexical |
| Lexical variation in <i>splinter</i> | splinter | spelk, speel | 10 | Lexical |
| Lexical variation in <i>snail</i> | snail | hodmedod, dod-man | 3 | Lexical |
| Pronunciation of <room> | room | [ʁʊm], [rʊm] | 3 | Phonolexical |
| Masculine reflexive pronoun | himself | himself, hisself | 2 | Morphological |
| Feminine possessive determiner | hers | hers, hern | 2 | Morphological |
| 3rd person habitual present | feed | do feed, feeds | 3 | Morphological |
| Velar Nasal Plus (cf. Wells 1982) – presence or absence | tongue | [tʌŋ], [tʌŋg] | 2 | Phonetic |
| Yod – presence or absence | new | [nju:], [nʊ:] | 2 | Phonetic |
| BATH vowel | last | [lɑ:t], [lɑst] | 3 | Phonetic |
| STRUT vowel | butter | ['bʊtə], ['bʌtə] | 2 | Phonetic |
| C/r/C realization (rhoticity) | arm | [ɑ:m], [ɑrm] | 2 | Phonetic |
| #/θ/C realization | three | [θri:], [fɹi:] | 4 | Phonetic |
| Intrusive /r/- presence or absence | thawing | [θɔ:ŋ], [θɔrŋ] | 2 | Phonetic |
| V/l/C realization | shelf | [ʃɛf], [ʃɛʊf] | 3 | Phonetic |
| KIT/SCHWA in unstressed syllables | pocket | ['pʰɔkɪt], ['pʰɔkət] | 2 | Phonetic |
| /ai/ before voiceless consonants | night | [nɛɪt], [nɪt] | 8 | Phonetic |
| /ai/ before voiced consonants | five | [fɛɪv], [fɔɪv] | 7 | Phonetic |
| Presence or absence of /h/ | hands | [handz] [andz] | 2 | Phonetic |
| CLOTH vowel | off | [ɔ:f], [ɑf] | 3 | Phonetic |
| MOUTH vowel | house | [hu:s], [hæus] | 8 | Phonetic |
| FACE vowel | bacon | ['bɪækən], ['be:kən] | 4 | Phonetic |
| V/t/V realization | bit of | [bɪd əv], [bɪʔ əv] | 4 | Phonetic |
| HAPPY vowel | happy | ['hæpi], ['hæpe] | 4 | Phonetic |
| Variation in <i>scone</i> | scone | [skaʊn], [skɒn] | 2 | Phonetic |
| Dative alternation | give it to me | give it me, give me it | 2 | Syntactic |

elements from earlier apps that were developed for Swiss German [26,29,30] and for varieties of German more generally [31]. The app features two core functionalities: (1) a dialect quiz and (2) dialect recordings. In what follows, we present the two functionalities along with the methods used for their implementation.

2.1. Dialect quiz

One of the core functionalities of EDA was to gather data on a wide number of linguistic variables – lexical, phonological and grammatical – via a quiz which predicted users' dialects based on their answers. The basis for the prediction was 25 discriminative maps of different linguistic variables from the SED [9]. To guess a user's dialect, we selected variables with distinct geographical distributions to help localize the quiz's prediction of the user's dialect as precisely as possible. Table 1 shows the 26 variables used, the prompt used for elicitation, example variants, the number of variants per variable, and the type of variable. The lexical variable 'splinter', for example, has ten different variants.

73% of the chosen variables are phonetic or phonolexical, 12% lexical, 12% morphological, and 3% syntactic. As for the number of variables selected, we assumed that too high a number of variables would leave users frustrated and many of them would opt out. At the same time, we expected dialect prediction to increase if there were more variables in the dialect quiz – to the point where prediction accuracy would be saturated. We intuitively felt that 25–30 variables would strike this balance, without applying statistical methods to arrive at this conclusion. Variables each showing different geographical distributions were chosen so that small areas could be distinguished from each other on the basis of a unique combination of variants across the set of variables. We did an overlay of 25 historical atlas maps, such as the maps for 'splinter' and 'butter', shown in Fig. 1.

These two variables alone divide England into North/South, while the North is further subdivided into individual pockets for different usages of 'splinter'. We added one variable which was not in the SED (and therefore could not be used for dialect prediction), the pronunciation of the vowel in 'scone'. The pronunciation of 'scone' was included because the variant pronunciations of it are, for English, highly variable, highly salient, and ideologically heavily loaded. Partly

because of this, we expected the inclusion of this variable to help spark media interest, critical for the widespread dissemination and uptake of the app. In practical terms, EDA prompts users to select their own pronunciation variant from a list by tapping on the smartphone screen. When variants cannot be written down because of only small phonetic differences (e.g. ['bʊtə] vs. ['bʌtə] for 'butter') the app shows phonetic transcriptions that are accompanied by audio recordings made by a native speaker of English (third author). Each variable is presented on a single screen with a lead phrase or triggering question at the top of the screen and variants listed below (Fig. 2, left and center), and the variables are presented in a randomized order for each user. Once users indicate which variants they use for all 26 variables the app presents a list of the three localities, out of a possible 313 adapted from the Basic Materials of the SED, that best correspond to their dialect and displays these on a map (Fig. 2, right). A heatmap shows the regions that correspond more or less strongly to the dialectal variants chosen by the user.

Underlying the dialect prediction is a table that contains a row for each locality, and a column for each pronunciation variant. Each cell features either a '1' (variant is present in the locality) or a '0' (variant is absent). For the columns (e.g. pronunciation variants) chosen by the user the 'algorithm' aggregates all the 1s row by row and identifies the row (locality) with the highest aggregation as the best hit (cf. [29]). If users believe the result to be accurate, they are shown a new screen informing them about how to support our research (Fig. 3, left). By clicking on 'OK, I'm in!', users comply with our privacy policy, which is explained on the screen. After clicking on this button, they are asked to indicate their dialect and a set of metadata. If they feel the first locality presented is not accurate, they can specify their dialect by moving a pin on a map to the location that best corresponds to their dialect (Fig. 3, center), before indicating metadata on age, gender, education, mobility, and ethnicity (Fig. 3, right). In both instances — correct or false dialect prediction — the location of a speaker's actual dialect is elicited. Users' pronunciation variants together with this metadata are anonymously stored on a server. None of the pieces of information elicited individually or in combination allow for identification of a user in the database. Users also can decline further input into our research, in which case they are shown the results screen again (Fig. 2, right).

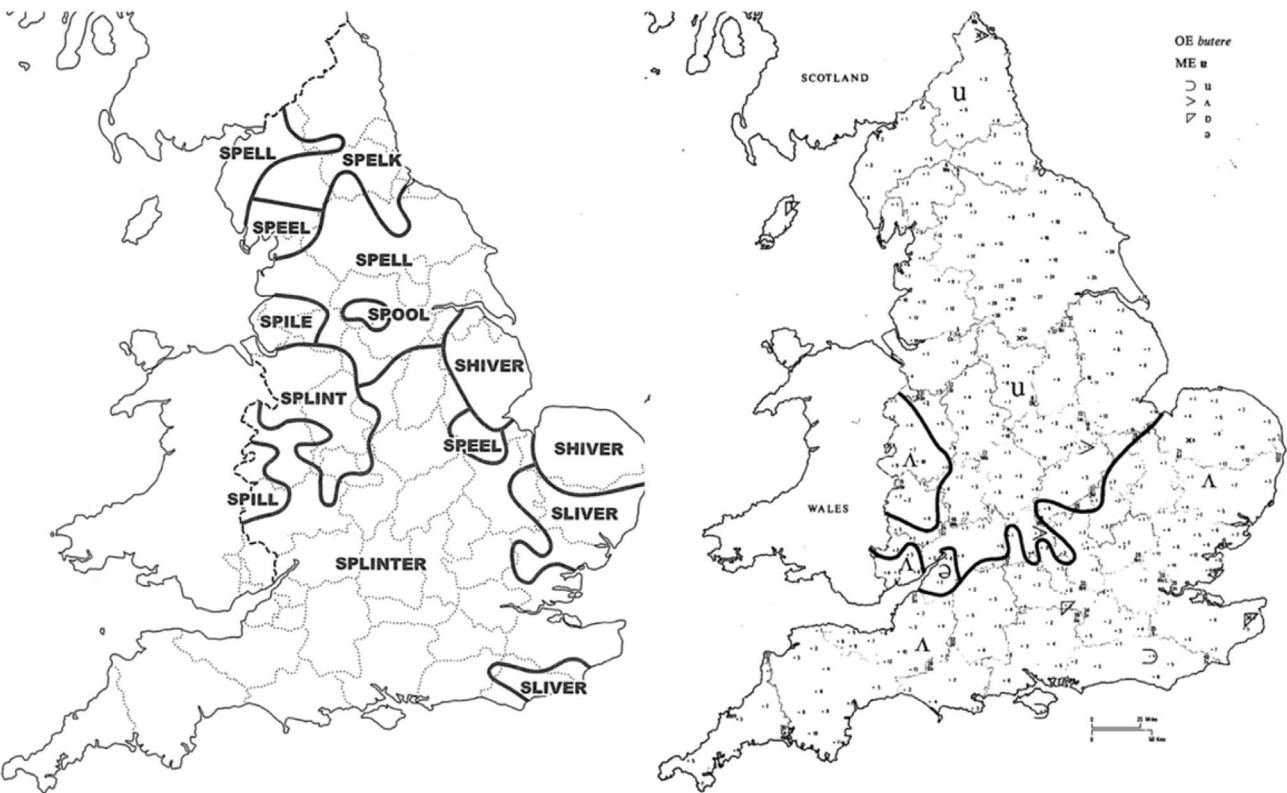


Fig. 1. ‘Splinter’ and ‘butter’ maps – two of the 25 maps used for dialect prediction in EDA (adopted from [9]).

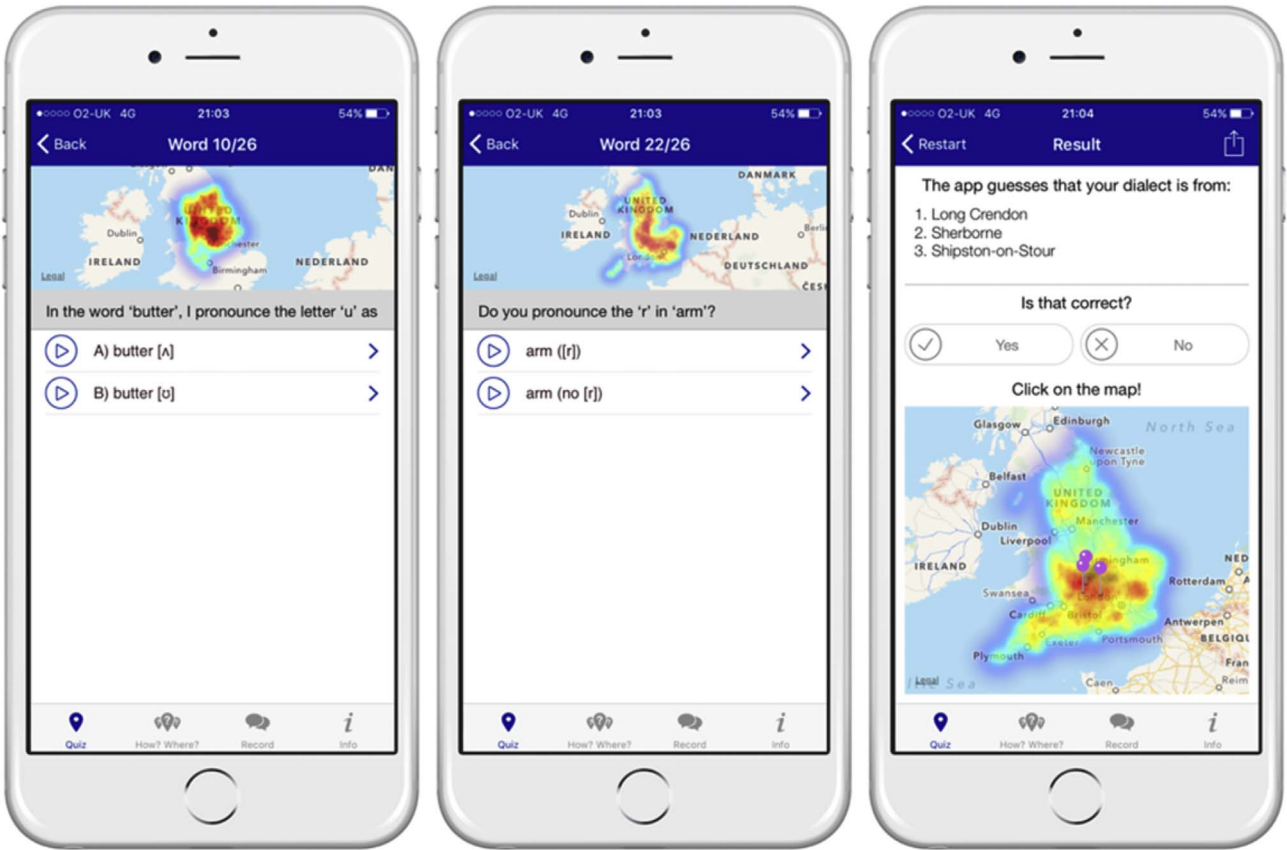


Fig. 2. Prompts for variant selection (left and center) and dialect quiz result (right).

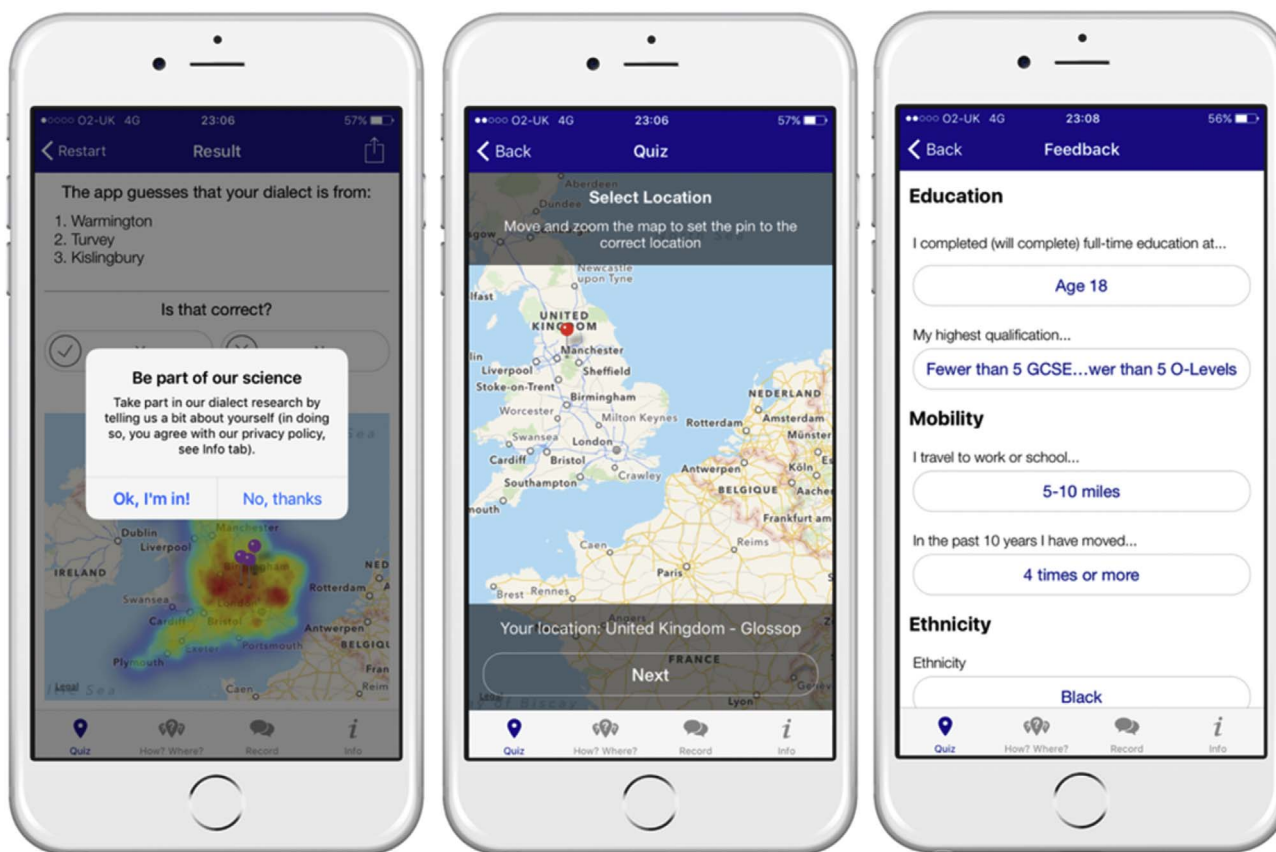


Fig. 3. Prompts for evaluating the dialect guessing result: consent form (left), placement of pin to locality that best corresponds to one's dialect (center), and prompt for metadata (right).

Collating users' answers to the quiz, their dialect location feedback and their metadata provides a contemporary socio-geographically differentiated snapshot of the dialect landscape of England, which can be used to investigate language change by comparing the app data with that from the SED.

2.2. Dialect recordings

The second core functionality allows users to anonymously record the readings of a short passage in their dialect, re-listen to these recordings, listen to recordings from the SED, and listen to the recordings uploaded by other app users by navigating an interactive map. The goal of this functionality was to provide users with dialect samples from all over England and therefore create awareness for dialectal variation, and to crowdsource speech of high recording quality, with little degradation due to the collection system – possible with today's smartphones. Previous research has shown that speech which is collected with a first-generation iPhone (i.e. in 2007) enables trustworthy measurements for formant analyses [32] and recordings done with contemporary smartphones even enable reliable analyses of voice quality in clinical settings [33]: today's smartphones typically process speech at 8 kHz–48 kHz sampling rates, 16-bit quantization rates, yet rates – particularly in the Android segment of the market – are device-dependent [34]. When it came to the selection of the sentence material for recordings, we considered a number of material design questions. Primarily, we were aiming for maximally broad phonetic coverage and material that was reasonably short and easy to read aloud (cf. material design considerations in [35]). We decided on 'The Boy who Cried Wolf' passage, see text in Fig. 4.

According to Deterding [36], this passage bears advantages over the standardly used 'The North Wind and the Sun', which has been criticized for lacking the occurrence of some sounds of English and because

of concerns regarding the measurement of speech rhythm. In addition, 'The Boy who Cried Wolf' is twice as long with less repetition of words. Here, the user interface prompts speakers first to indicate, i.e. self-declare, their dialect, by again placing the pin on the locality that best corresponds to their dialect (Fig. 3, center) and to fill in the metadata described in 2.1 (Fig. 3, right panel). On the metadata screen it is explained that in recording their voices, the users consent to our privacy statement. Again, none of these pieces of information individually or in combination allow for the identification of a user. Users also have the opportunity to opt out of this procedure at any time (i.e. by clicking on 'back' in the top left corner in the screens shown in Fig. 5). They then proceed to the recording instructions, which read 'Please record your voice in a quiet place. Hold your device approximately 6 inches/15 cm from your mouth. Please use your regional accent or dialect and speak in the way you would talk to your friends from home'. Users then record the sentences shown on the screen. The passage 'The Boy who Cried Wolf' consists of ten sentences, which are all shown on individual screens, see Fig. 5 left.

They click on a record button on the screen and read the sentence out loud. Once the sentence has been read, they click on the stop button. They can subsequently click 'Play' to play back their recording (see Fig. 5 left panel). If they are pleased with the current recording they click continue to proceed to the next sentence (see Fig. 5 left). If they dislike their recording, they click 'Record' again to re-record the sentence, which they can do as many times as they like (cf. [37]). Recordings are anonymously uploaded to servers where each audio file is given a unique ID. Once their recording has been uploaded, users can navigate to an interactive map (Fig. 5, center) where they can listen to their own recordings, those of other users, and historical recordings from the Survey of English Dialects (Fig. 5, right). The 'How? Where?' tab visible at the bottom of the screens in Fig. 5 allows users to select localities and discover their local dialect according to the SED – or to

There was once a poor shepherd boy who used to watch his flocks in the fields next to a dark forest near the foot of a mountain. One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun. Raising his fist in the air, he ran down to the village shouting ‘Wolf, Wolf.’ As soon as they heard him, the villagers all rushed from their homes, full of concern for his safety, and two of his cousins even stayed with him for a short while. This gave the boy so much pleasure that a few days later he tried exactly the same trick again, and once more he was successful. However, not long after, a wolf that had just escaped from the zoo was looking for a change from its usual diet of chicken and duck. So, overcoming its fear of being shot, it actually did come out from the forest and began to threaten the sheep. Racing down to the village, the boy of course cried out even louder than before. Unfortunately, as all the villagers were convinced that he was trying to fool them a third time, they told him, ‘Go away and don’t bother us again.’ And so the wolf had a feast.

Fig. 4. ‘The Boy who Cried Wolf’ passage (adopted from [36]).

select dialectal variants and to discover where these were in use, also according to the SED. And the info tab provides some information about the app’s functions, its methods, its developers, its privacy policy, and presents acknowledgements.

2.3. Technical app development

EDA was developed for iOS and Android in Xcode using Objective-C and is available for free download in the respective app stores. Dialect quiz data, quiz and acoustic-phonetic metadata elicited by EDA are stored in a relational MySQL database (InnoDB), audio files are stored in wave format on a server and are cross-referenced in the database. An application programming interface (API) enables the communication between the mobile application and the database, and a backend platform enables the administration of the data using CakePHP and Bootstrap (cf. [29]). The app was developed over months of extensive pilot testing and is based on previous experiences made during app

development, i.e. testing for human-machine interaction, the comprehensibility of the on-screen instructions, the types of prompts, the order of presentation of prompts etc. to insure smooth uptake by participants.

3. English Dialects App corpus (EDAC) statistics

To recruit a maximum number of participants, we worked closely with the University of Cambridge and University of Bern press offices. The app was announced in press releases, newspapers, radio and television interviews. Data collection in earnest began with the release of the app in January 2016 and is ongoing. In mid-May of 2016 we released an update of the app for iOS and Android, in which we (a) improved dialect prediction using app-collected data from the dialect quiz over the first four months, (b) added a question about multiple submission, users’ parents’ educational backgrounds, and ‘other’ as an option in gender in the metadata, and (c) performed minor bug fixes. More than 99,000 people had downloaded the app by May 2017. During this



Fig. 5. Prompts for dialect recordings: recording of sentence (left), map displaying localities with recordings (center), and overview of recordings from one locality (right).

1.5-year cycle of operation, the EDAC protocol produced more data at faster rates than has been collected in English dialectology before. So far, the app has collected dialect quiz data from more than 50,700 users and audio data from more than 4300 speakers.

One concern in corpus-based dialectology is sampling (cf. [4]): typically, corpus-based dialectological approaches aim to cross-compare contemporary data to historical data. This historical data was normally sampled in a different way, e.g. NORMs (see introduction). Modern-day approaches – like EDA – sample subsets of entire, heterogeneous populations, of which NORMs are only a small fragment. This makes cross-comparisons between the different corpora difficult as the sampling techniques are vastly different. With EDA we sought to collect data from a balanced sample of the UK population; however, this was understandably not possible, given the demographics of smartphone ownership and engagement with mobile technologies. EDA, therefore, perhaps not surprisingly, oversampled young adults. Also, the app was mainly advertised through the public media (e.g. *BBC*, *DailyMail*, *The Telegraph*, *Reddit*, *phys. org* etc., cf. <https://sites.google.com/site/adrianleemann/press>), its users are therefore additionally biased towards the demographic of public media consumption. The corpus statistics presented below try to address such sampling issues and contextualize the descriptive statistics in relation to the population of the UK in general.

3.1. Dialect quiz database statistics

By May 2017, 50,755 users from 128 countries had evaluated the quiz result; 1262 users who participated in the second cycle submitted the information more than once and were thus excluded from the statistics presented below. In what follows, we present descriptive statistics of users stemming from the UK, the Channel Islands, Isle of Man, and the Republic of Ireland, which amount to 47,059 in total – henceforth “the corpus”.

3.1.1. Geographical distribution

As shown in Fig. 3 (center), users were asked to place a pin on the locality that best corresponds to the dialect they speak. Fig. 6 shows all localities that are represented in the corpus (left) as well as the density of these localities with a heatmap (center), and the population density

of the UK and the Republic of Ireland as of 2011 (right) (adopted from the European Forum for Geostatistics, <http://i.imgur.com/jvhxb5L.jpg>).

These latitudes/longitudes belong to 4921 villages or cities. As can be seen from the leftmost panel, much of England is represented in the corpus; there is, understandably, less data from very sparsely populated parts of England. The density map (center) reveals that the highest density of responses is found in the Southeast as well as in the Northwestern Midlands. This closely mirrors the general population density of the UK and the Republic of Ireland shown in the right-most panel in Fig. 6. Users from the Republic of Ireland, Northern Ireland, and Scotland are understandably under-represented – the app was specifically targeted at users from England, cf. Section 1.

3.1.2. Age and gender

The mean age of the 47,059 users is 34.9 (SD = 14.3) – the median is 32. 48% (N = 22,572) are females, 52% (N = 24,423) are males, and 0.13% indicated ‘other’ (N = 64); this latter category was added only in the second cycle of data collection. Fig. 7 plots age counts as a function of age: males (left), females (right). The blue line in the left panel indicates the actual distribution for males, the red line shows the expected distribution if we had done a perfect sampling according to the age distributions found by the Office for National Statistics (ONS) for mid-2015 (cf. [38]. for England, Wales, Scotland, and Northern Ireland). The yellow line in the right panel indicates the actual distribution for females, the green line the expected distribution if we had done a perfect sampling of age in the population (cf. [38]).

As can be seen from Fig. 7, for both sexes, speakers < 14 years of age as well as age groups > 52 are undersampled; at the same time, people between 14 and 52, especially so between 20 and 32, are oversampled for both sexes. This relationship can partly be explained by smartphone ownership penetration in the UK in 2015: Statista [39] reports that 90% of people between 16 and 24 own a smartphone, and 87% of 25–34 year olds, 80% of 35–54 year olds, but only 50% of 55–64 year olds and only 18% of 65 + year olds have a smartphone. Using a smartphone app to generate a dialect corpus therefore targets a specific segment of the population, which might help explain this skewness.

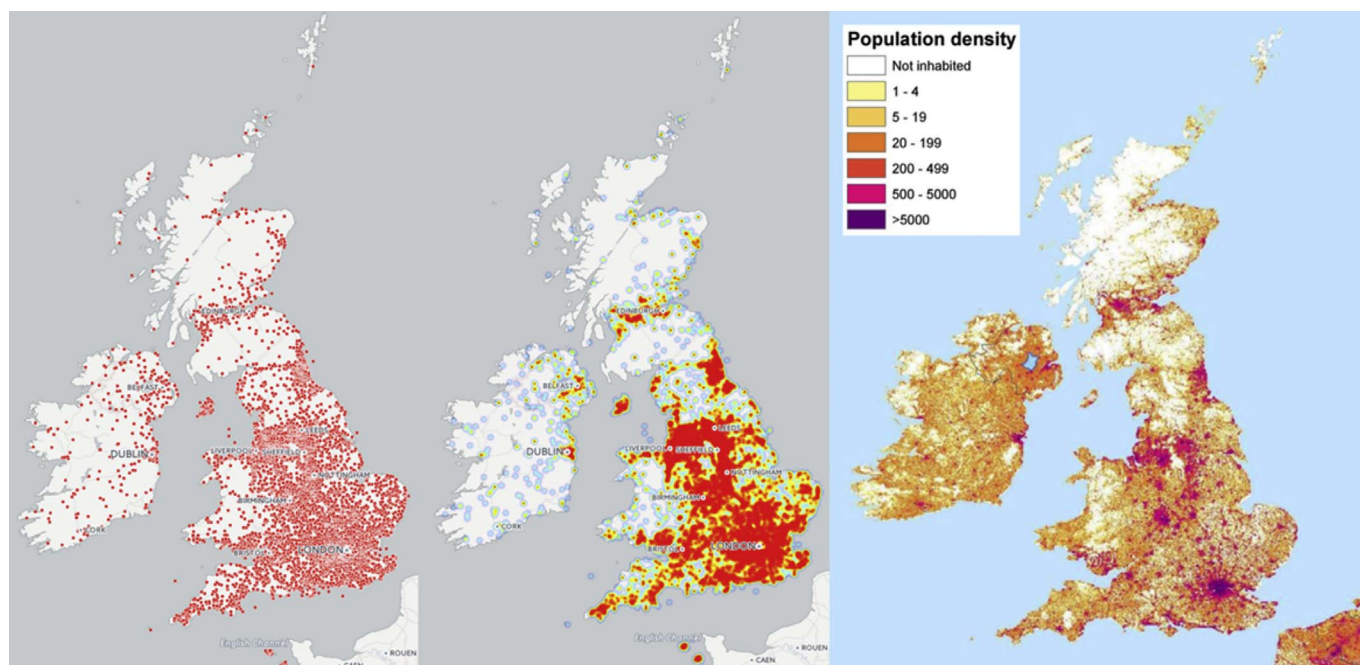


Fig. 6. Geographical distribution of dialect quiz users: unique localities (left), density of localities (center), UK population and Republic of Ireland population density (right).

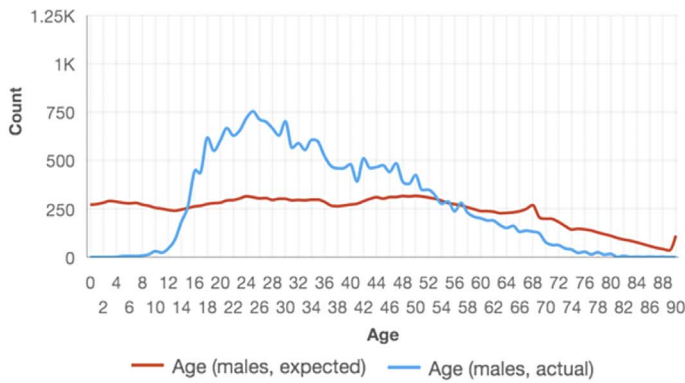


Fig. 7. Age distribution of dialect quiz users.

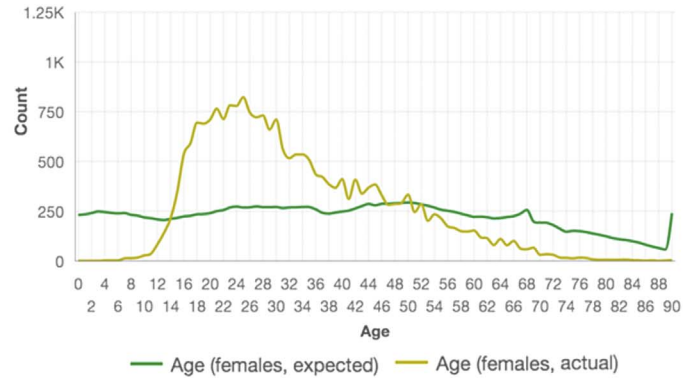


Fig. 8. Ethnic distribution, expected (yellow) and actual (purple).

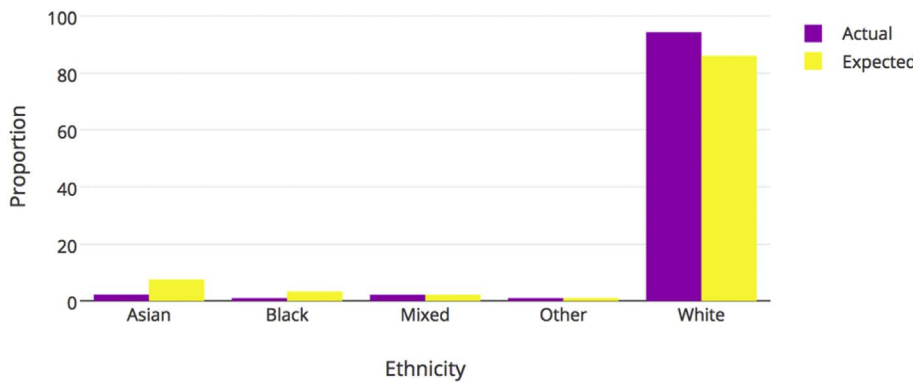


Fig. 9. Qualification distribution, expected (orange) and actual (blue).

3.1.3. Ethnicity

Further descriptive statistics revealed that in terms of ethnicity, our app-collected data does not entirely reflect the distribution found in the population. Fig. 8 plots the proportions of the major ethnicity categories: the yellow bars show the expected distributions if our sample was balanced by real population according to Census 2011 (cf. [40]. for England and Wales); the purple bars show the actual distribution in our corpus.

In Census 2011, 86% reported to be of White ethnicity, in our sample this proportion is almost 10% more, 94.3%. On the other hand, people of Asian (−5.3%) and Black ethnicity (−2.7%) are under-sampled. People with mixed and other ethnicities are reflected quite similarly as in the English and Welsh population. The app further asked for precise ethnicity. For White ethnicity, 87.7% indicated White English, 4.1% Other White, 2.7% White Scottish, 2.6% White Welsh, 1.5% White Irish, 0.8% White Northern Irish, and 0.07% Gypsy or Irish Traveler.

3.1.4. Qualification and education

Users further provided metadata on their highest qualification. Fig. 9 shows the proportions of the highest qualification categories: the orange bars indicate the expected distribution if our sample was balanced by real population according to Census 2011 (cf. [41]. for England and Wales); the blue bars show the actual distribution in our sample. Qualification levels were grouped according to indicators provided at ONS [42].

Fig. 9 reveals a substantial oversampling of people with higher educational qualifications in EDAC: 29.7% in Census 2011 vs. 67.4% in the app. All other categories – except for speakers with A levels who are also oversampled (+4.3%) – are under-sampled; especially so people with no qualifications or fewer than 5 GCSEs or equivalent (a GCSE is a school qualification in a specific subject, such as English Literature or Math or Geography that is usually taken by English school pupils during the school year in which they turn 16 years old. GCSEs were preceded historically by ‘(Ordinary)-Levels’ and CSEs). It could be argued that the topics touched upon by EDA – linguistics, dialectology, phonetics – are

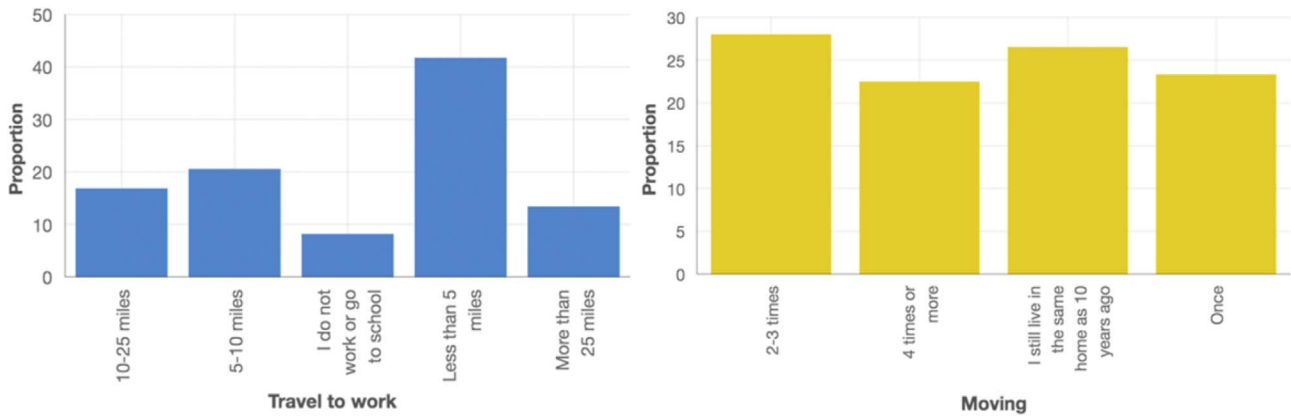


Fig. 10. Mobility distribution of users in quiz data.

likely to attract a more educated audience in England, which results in the bias towards a highly-educated sample. The same may hold for smartphone penetration and media consumption. We further asked users to provide information as to the age of completion of their education. 71.4% reported that they completed their education at or over the age of 19, 14.8% at age 18, 11.3% at age 16–17, and 2.5% at age 15 or under.

3.1.5. Mobility

To compare linguistic behavior between speakers who are mobile vs. those who are non-mobile, EDA elicited data on how far people travel to work and on how many times they have moved in the past ten years. Fig. 10 shows the breakdown of the answers to the question ‘I travel to work or school ...’ (left hand-side) and ‘In the past 10 years I have moved ...’ (right hand-side).

42% of the users who provided feedback travel to work less than five miles; nearly a quarter (20.4%) travels 5–10 miles, 17% travel 10–25 miles and 13.2% more than 25 miles. 7.4% do not work or go to school. According to statistics reported in Census 2011 [43], 81% of English and Welsh citizens undertake a regular commute to work, with an average travel distance of 15 km, i.e. around 10 miles. This distance is heavily dependent on the region, with people from the East of England traveling the most (17 km) and those from London the least (11 km). Males further tend to commute farther than females and workers between 35 and 39 travel the farthest [43]. Earlier research by Champion [44] showed that around 48% of working age residents either do not commute at all, or commute less than 5 km, and about 12% commute more than 20 km. There are wide regional differences,

however, within these figures, with rural residents much more likely to commute longer distances ([44]; p. 171). The app sample for commuting appears to be not too dissimilar to that found in the census data. Regarding moving places, more than a quarter of the users (28%) moved 2–3 times in the past ten years; 22% even moved more than 4 times. Virtually half still lives in the same place as ten years ago or has moved only once. These figures appear to reflect general moving trends in the UK: ONS data from Census 2011 suggest that 11.4% of the population living in the UK had moved in the last year [45], with most moves being over a short distance. Most movers (59.6%) had stayed within the same local authority district. In the census data, there was also an effect of sex, with males being slightly more migratory, and an effect of age, with younger adults showing a higher propensity to move.

3.2. Acoustic-phonetic database statistics

In total, 4310 users from 91 countries provided audio recordings. In what follows, we only present descriptive statistics from users stemming from the UK, Guernsey, Jersey, Isle of Man, and the Republic of Ireland, which amount to 3551 speakers in total. The corpus is composed of ten sentences, making up the ‘The Boy who Cried Wolf’ passage. Virtually all speakers recorded all the ten sentences. A few users did not record the sentences that appeared later in the passage. The first sentence thus has the most recordings ($N = 3495$), the last sentence – sentence 10 – the least recordings ($N = 3262$).

3.2.1. Geographical distribution

When users conduct the audio recordings, they are asked to place a



Fig. 11. Localities represented in the acoustic-phonetic corpus (left), geographical distribution of speakers (center), UK population and Republic of Ireland population density.

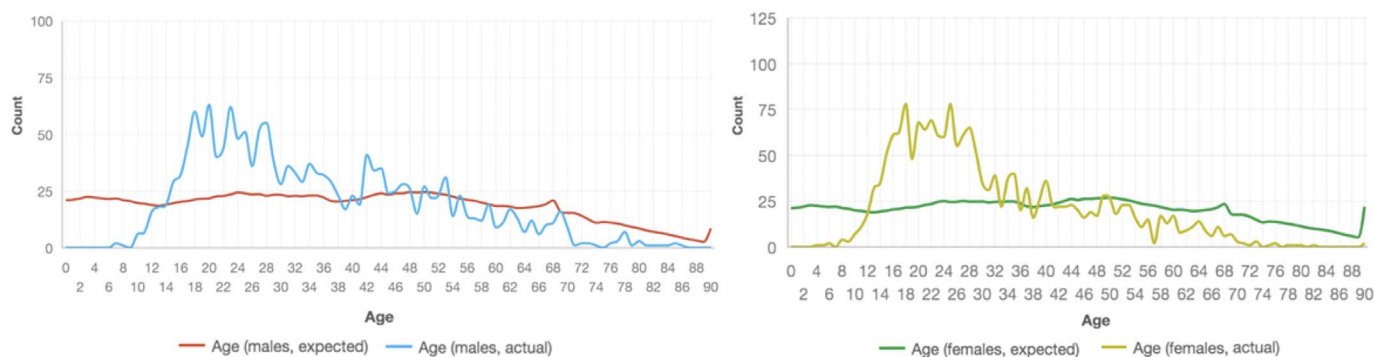


Fig. 12. Age distribution of speakers in the acoustic-phonetic corpus.

pin on the locality that best corresponds to the dialect they speak (see Fig. 3, center). If users do the quiz before the recording, this information is filled in automatically for the audio recording functionality. Users identified 1336 unique latitudes and longitudes, shown in Fig. 11 (left). The center panel in Fig. 11 shows a clustered display of the number of speakers and the right panel shows overall UK and the Republic of Ireland population density as of 2011 (right) (adopted from the European Forum for Geostatistics, <http://i.imgur.com/jvhxb5L.jpg>).

The cluster map (center) reveals that the highest density of responses is found in the Southeast, as well as in the Northwestern Midlands – thus reflecting the general population density of England shown in the right-most panel in Fig. 11.

3.2.2. Age and gender

The mean age of the 3551 speakers is 33.3 (SD = 15.5); the median is 29. A little more than half of the users, 52.2% (N = 1853), are females, 47.6% (N = 1689) are males, and 0.2% indicated ‘other’ (N = 9); this latter category was added only when an updated version of the app was released, however. Fig. 12 plots age counts as a function of age: males (left), females (right). The blue line in the left panel indicates the actual distribution for males, the red line the expected distribution if it was balanced by real population according to mid-2015 (cf. [38], for UK, England and Wales, Scotland, and Northern Ireland). The yellow line in the right panel indicates the actual distribution for females, the green line the expected distribution if it was balanced by real population (cf. [38]).

As can be seen from Fig. 12, speakers < 14 years of age are undersampled. For males (left panel) speakers 56 and above are undersampled; for females (right panel) speakers 44 years of age and above are undersampled. For both sexes, however, speakers between 14 and 28 are oversampled. Van Leeuwen et al. [46] report highly similar over- and under-sampling trends on crowdsourcing speakers of Dutch with their app *Sprekend Nederland*. As was argued in the age and gender distribution discussion for the quiz data (cf. Fig. 7), this relationship between actual and expected distributions of speaker age can probably be explained by smartphone ownership penetration in the UK in 2015.

3.2.3. Ethnicity

As for ethnic distributions in the acoustic-phonetic corpus, EDA sampled data that reflected UK distributions more accurately than in the dialect quiz corpus. Fig. 13 plots the proportional ethnicity categories: the yellow bars indicate the expected distribution if our sample was structured according to Census 2011 (cf. [40], for England and Wales); the purple bars show the actual distribution in the app.

87.5% of the sampled population is White whereas in Census 2011 this category accounted for 86% of the population. EDA somewhat under-sampled speakers of Asian descent (–2.3%) and of Black descent (–1.2%). Speakers of mixed descent are over-sampled slightly (+1.6%) and those of other descent largely reflect population

distributions. For White ethnicity, 84% indicated White English, 6% other White, 3% White Scottish, 3% White Irish, 2.7% White Welsh, 1.3% White Northern Irish, and 0.2% indicated Gypsy or Irish Traveler.

3.2.4. Qualification and education

Fig. 14 shows the proportions of highest qualifications found in the acoustic-phonetic corpus: the orange bars show the expected distribution if the sample was distributed according to Census 2011 (cf. [41], for England and Wales); the blue bars show the actual distribution. Qualification levels were grouped according to indicators provided by ONS [42].

Fig. 14 shows that, here too, speakers with higher educational qualifications are oversampled (29.7% Census 2011 vs. 57% EDA acoustic-phonetic corpus), albeit the discrepancy between the expected and actual distribution is somewhat less substantial than in the quiz corpus, where 67.4% indicated to have a higher education degree. All other categories – except for speakers with A levels who are also over-sampled (+6%) – are under-sampled, particularly so people with no qualifications or fewer than 5 GCSEs. Descriptive statistics further revealed that 65% of the speakers in the EDA acoustic-phonetic corpus finished their education over the age of 19, 17% at 18, 13.5% at 16–17, and 4.5% at age 15 or under.

3.2.5. Mobility

Fig. 15 shows the breakdown of the answers to the question ‘I travel to work or school ...’ (left hand-side) and ‘In the past 10 years I have moved ...’ (right hand-side).

41% of the speakers travel to work less than five miles; nearly a quarter (22%) travels 5–10 miles, 15% travel 10–25 miles and 13% more than 25 miles. 9% do not work or go to school. Compared to ONS statistics for England and Wales [43] and Champion [44] presented in Section 3.1.5, this sample, too, appears to be relatively well-balanced. With regard to moving places, more than a quarter of the users (29%) moved 2–3 times in the past ten years; 22% even moved more than 4 times. Yet, half (49%) still lives in the same place as ten years ago or has moved only once. The figures here reflect general moving trends in the UK, presented in the quiz data section [45].

Overall, then, the sample is biased in favor of a certain young, digitally-native, highly educated demographic. Without further differentiation of the sample, this could be seriously problematic for the claims we might want to make about the dialect regions of the country. We collected, as mentioned earlier, however, systematic social meta-data from users, enabling us to compare different demographic groups, extract data only for certain groups if desired, and therefore appropriately manage imbalances within the sample as a whole.

4. Use for theoretical and applied research

The development of a contemporary dialect and acoustic-phonetic corpus of English dialects spoken in the UK is of primary importance, of

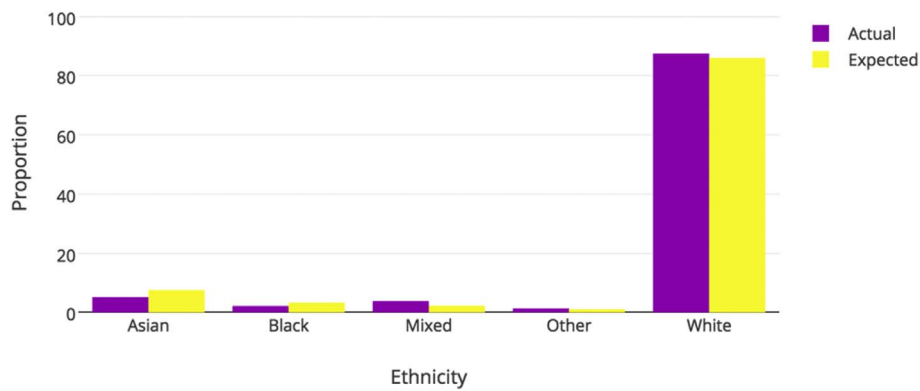


Fig. 13. Ethnicity distribution, expected (yellow) and actual (purple).

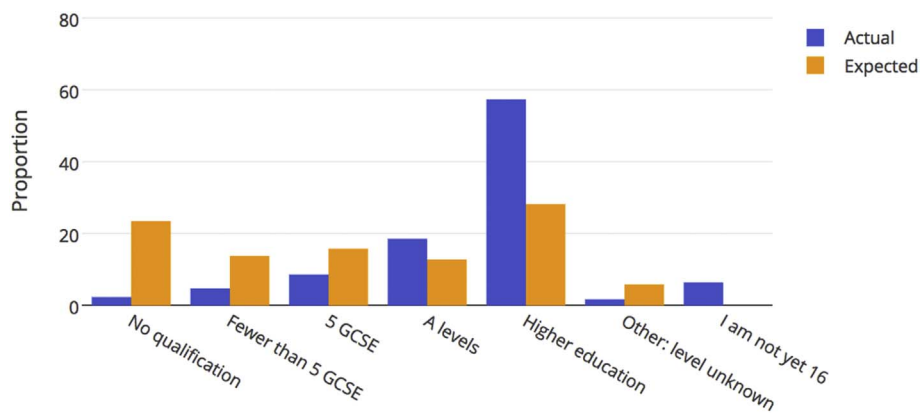


Fig. 14. Qualification distribution, expected (orange) and actual (blue).

course, for dialectology, but also significant – on a more applied level – for forensic phonetics.

4.1. Use for theoretical research

4.1.1. Prediction corpus

Most importantly, the quiz corpus allows for analyses of language change. When the users evaluate the quiz result (cf. 2.1), we elicit the speakers' dialects and the 26 variants they indicated, which can then be compared to the variants indicated for these localities in the SED. To showcase this procedure, Fig. 16 provides preliminary maps of our results that we generated for a press release. The app question underlying this variable was 'I pronounce the word 'arm' as (a) arm ([r]) or arm (no

[r])'. The left panel shows the distribution of the two variants found in the SED, the right panel shows that found in EDA's quiz corpus, based on answers from roughly 29,000 respondents. The greener the area, the less the [r] is pronounced; the redder the area, the more the [r] is pronounced.

The map was created by Tam Blaxter (University of Cambridge) in QGIS [48], applying a normalization that averages values across the 100 nearest neighbors for each locality [49]. This map is quite revealing in several ways: (a) we observe a general trend that the non-rhotic forms (green, no [r]) are considerably more widespread in 2016 than in the 1950s (at the time the SED was conducted), (b) unlike in the SED, the EDA also (unintentionally it must be said) gathered data from Wales, Scotland, Northern Ireland, the Republic of Ireland, and the

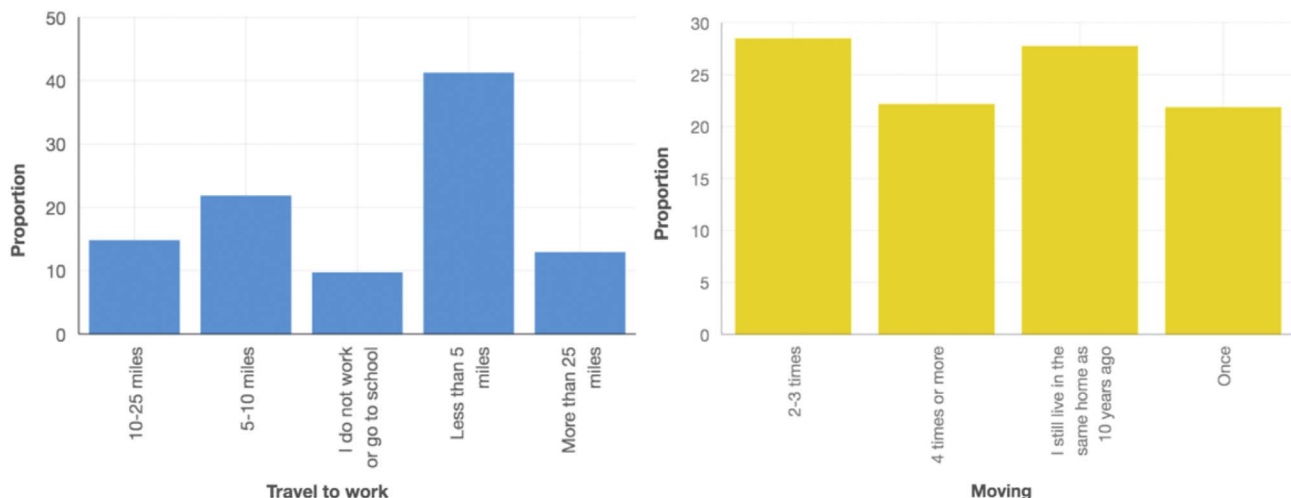


Fig. 15. Mobility distribution of speakers in the acoustic-phonetic corpus.

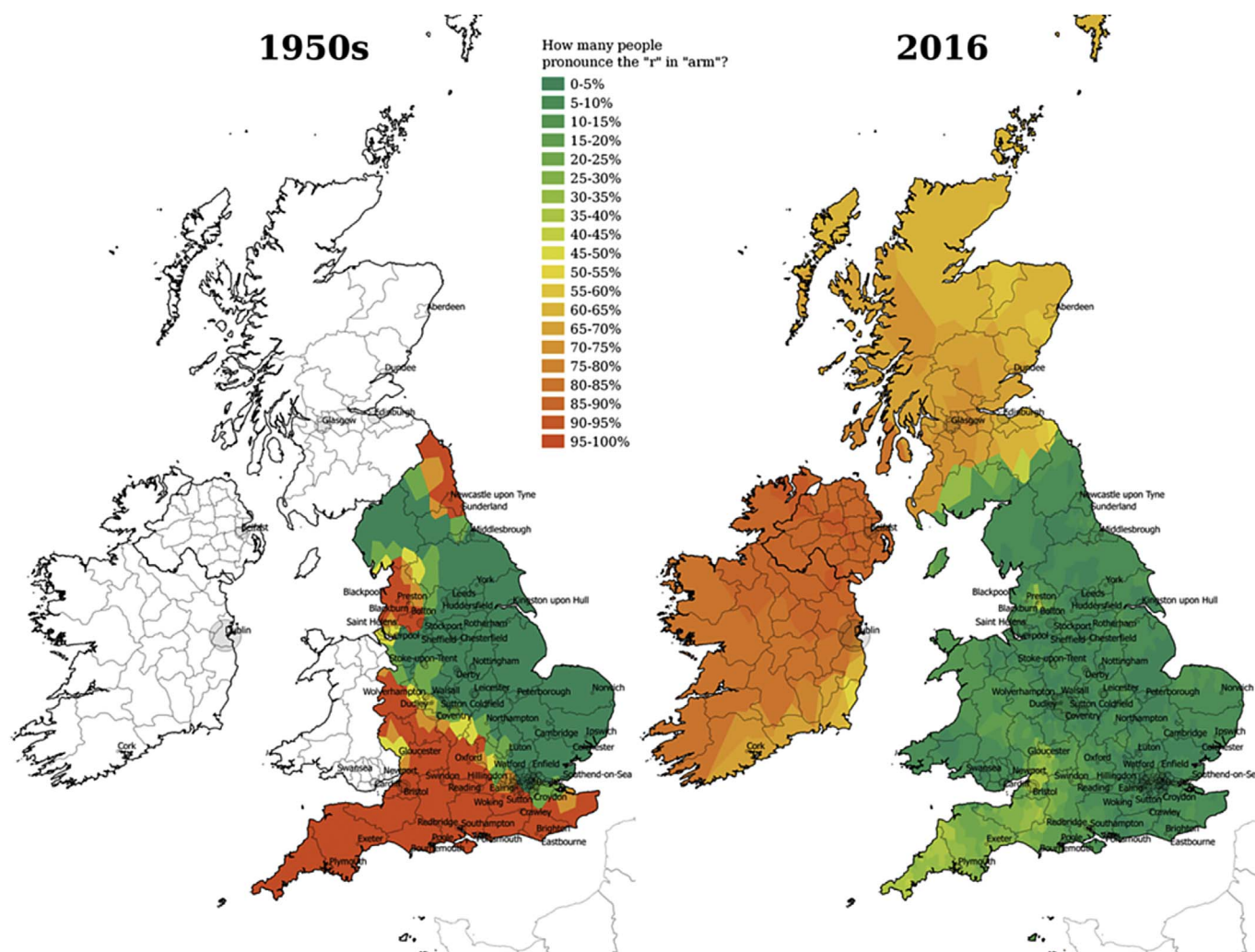


Fig. 16. Geographical distribution of rhoticity based on the SED data (left) and EDA quiz data (right) (graphic by Tam Blaxter (University of Cambridge) – adapted from [47]).

Channel Islands, (c) the high spatial resolution of the EDAC enables a highly fine-grained analysis of language change, e.g. there are places in the South-West of England (especially urban centers such as Exeter, Plymouth, Gloucester and, especially, Bristol) which still show rhoticity to varying degrees, reflecting remnants of the 1950s variants. Indeed, Grossenbacher [50], investigating the presence or absence of rhoticity in both parts of the corpus – the quiz and the acoustic-phonetic data – in the area around Bristol, demonstrated that in both there was a shift away from rhoticity the younger the users were, but also a shift away from rhoticity in rural areas. Among the young, then, rhoticity was a markedly *urban* characteristic.

More generally, then, by comparing data retrieved through the quiz function to historical data, we can investigate changes in the regional distributions of dialect features, track the regional spread of innovations and address, on the basis of nationwide data, questions such as whether leveling – the loss of minority dialect forms and regional convergence towards majority features – has affected all dialects similarly, or whether some dialect areas are more strongly affected than others. Does the degree of leveling found in different regions pattern with differences in urbanization, counter-urbanization, and migration? Are regional centers implicated in leveling towards emerging local norms? Where, today, are the relic areas – those places where obsolescent dialect variants cling on to life? The data will also highlight particular localities and speech communities deserving of further, more detailed, ethnographically sensitive studies. The fact that recent

quantitative work still makes use of the 1950s SED data (e.g. [13,51]) points towards the value of EDAC for contemporary English dialectology.

4.1.2. Acoustic-phonetic corpus

The amount of material and the number of speakers collected in the acoustic-phonetic corpus will make it possible to run a number of important experiments. At first, the acoustic-phonetic data needs to be transcribed and tagged. A purely manual phonetic transcription, while feasible in principle, is prohibitively labor-intensive to generate – which is where forced aligners come to use. The forced alignment of the audio data is currently under way. Each sentence is being forced aligned using MAUS [52], allowing for a time-aligned segmentation. MAUS simply needs to be fed a wave file and a corresponding text file that contains the sentence material that was read. Fig. 17 shows an example of a MAUS-aligned signal: the first phrase of sentence 4 of ‘The Boy who Cried Wolf’ as articulated by a speaker from Cambridge in our acoustic-phonetic corpus.

Tier one shows the word-by-word orthographic transcription, tier two the word-by-word SAMPA transcription, and tier three the phone-by-phone SAMPA transcription. The blue line in the spectrogram demonstrates the tracked intonation contour. As Fig. 17 shows, force aligning crowdsourced audio data is very possible – segment onsets and offglides are aligned quite accurately, without having yet applied any manual correction on the data. This type of time-aligned orthographic

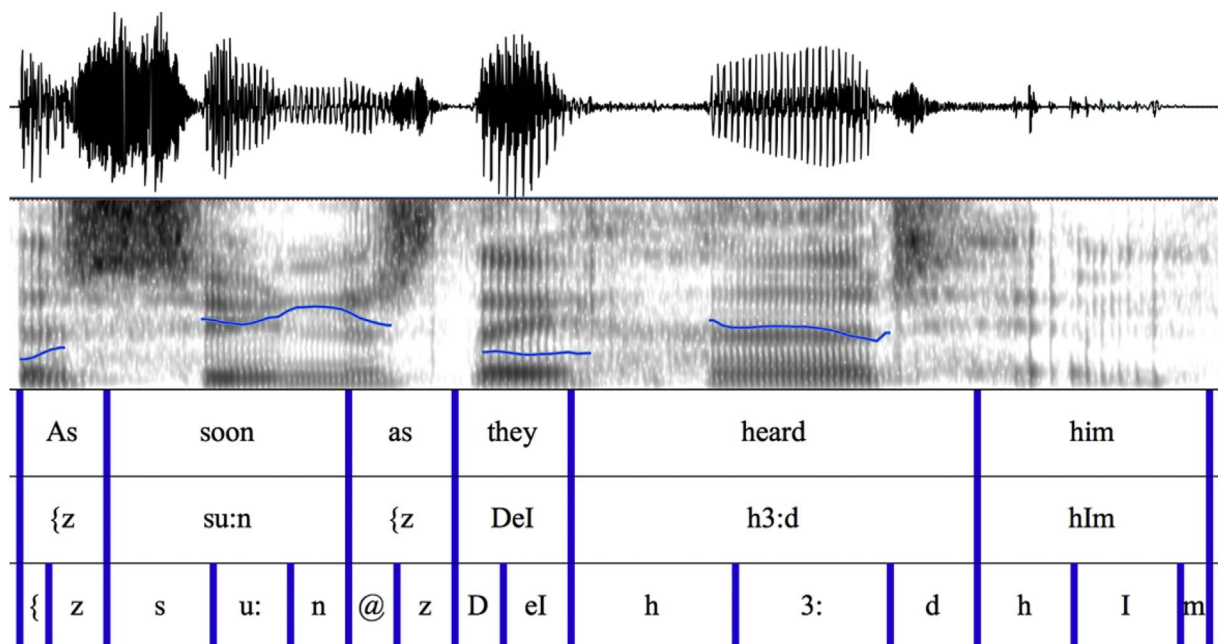


Fig. 17. Forced aligned file from EDA's acoustic-phonetic corpus.

transcription is useful when searching for a specific word, while the time-aligned phonetic transcriptions can be used to relate the lexical representations of words to their acoustic, segment-by-segment realizations. The corpus will be annotated manually – e.g. for stress, word class, intonation etc. This annotation will enable synchronic, cross-dialectal comparison of segments and prosody. Further, it will enable diachronic acoustic analyses of prosodic features comparing contemporary to historic data for localities for which there are historical SED recordings. This is possible because suprasegmental patterns typically remain intact even in historical, often distorted recordings. Such real-time, diachronic acoustic analyses of dialects are currently largely non-existent, with only few exceptions (cf. [53]).

A simultaneous examination of the two corpora (4.1.1 and 4.1.2) will reveal new insights about lay people's intuitions about dialect use. Native intuitions concerning grammaticality provide the central core of data for much theoretical linguistics, a fact which has often been criticized both from within the field and by sociolinguists outside it on the grounds that native intuitions do not consistently reflect use (cf. [54]). Yet the studies of this mismatch between intuition and use are all relatively localized and small scale. To date, no large scale survey considering a large set of dialect features with a sufficient number of participants across many speech communities has been conducted to quantify the problem of unreliable intuitions. The users who fill out the dialect quiz – essentially a dialect intuition task – and make audio recordings – a dialect production task – provide an unparalleled dataset for comparing intuitions and use: by comparing the answers users give in the quiz to their actual pronunciation of those words in the recording, researchers will be able to identify where speakers' self-belief concerning pronunciation is and is not accurate. Existing studies based on online language survey data (such as [55]) are vulnerable to the criticism that they are unable to quantify the error introduced by under- and over-reporting of socially marked variants. Quantifying rates of under- and over-reporting in the matched acoustic-phonetic and quiz corpora will enable researchers to be confident in the proper interpretation of results from the quiz corpus (cf. [56]).

4.2. Use for applied research

The dialect quiz and the acoustic-phonetic corpus both bear relevance for forensic phonetics. In forensic casework, particularly in

forensic speaker comparison, experts compare speech in criminal and suspect recordings to evaluate whether a criminal's voice is that of a suspect or of someone else. Findings of a survey on forensic speaker comparison practices found that 28% of practitioners examine dialect and accentual features [57]. Experts use different strategies to identify a speaker's regional background: they may perform auditory and acoustic analyses and/or consult audio databases and literature on dialects [58]. This consultation of literature also includes the consultation of dialect Atlases. As stated earlier, much of the area-covering material on regional variation on accents of English is outdated; dialect isoglosses have shifted considerably in the meantime. With our analyses we provide an updated picture for regional variation in the UK, which provides forensic experts with a comprehensive database of unprecedented spatial resolution.

The app's acoustic-phonetic audio corpus, too, bears relevance for forensic casework. To give an estimation of the probability that a speaker in both the case and the suspect material are the same, experts look for high similarity in the speech signals on the basis of analyzing various acoustic parameters (e.g. f0, speaking rate, formant frequencies, etc.) [59]. At the same time, when comparing the speakers, experts particularly look for low typicality manifestations of features in the speech signals, i.e. manifestations of acoustic features that are not very common in the population. As an example, if known and disputed samples both exhibit extremely fast articulation rates – which would be a manifestation of a low typicality feature when looking at a general population, i.e. extremely fast articulation rates do not occur very often – this would increase the likelihood that the samples come from the same speakers. If, on the other hand, both speakers articulate with average speed (not particularly fast or slow), i.e. which is a manifestation of speaking rate with high typicality in the general population, then speaking rate as a diagnostic to identify speakers would not be very helpful – as most people in the population exhibit average speaking rates. To know if a manifestation of a feature exhibits high or low typicality, forensic practitioners need to know population distributions of these features (cf. [60]). Aside from a few global speech characteristics (such as f0 for example, see below), there is a lack of information on the distribution of speech features across the population. In a pilot study using EDAC, Hudson et al. [61] analyzed coarse f0 statistics on a preliminary 2000 speaker set of the acoustic-phonetic corpus. Fig. 18 shows mean f0 for approximately 800 males.

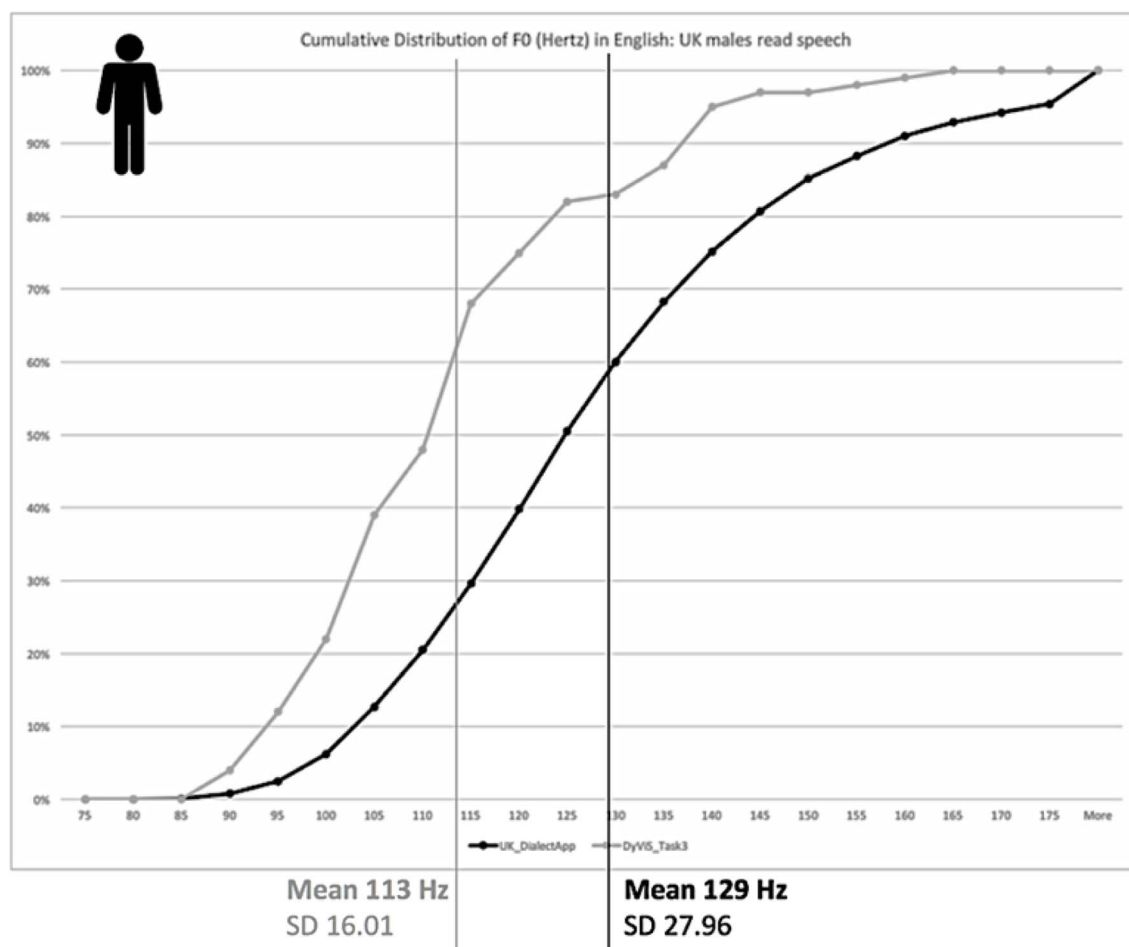


Fig. 18. Provisional population distribution of mean f0 based on EDA's acoustic -phonetic corpus (adopted from [61]).

For male adults, they report a mean f0 of 129 Hz with a standard deviation of 27.96. As a point of reference, they present the DyViS database ('Task 3') [62] in Fig. 18 (grey line), which reports a mean f0 of 113 Hz (SD 16.01) for educated SSBE speakers. We observe a higher f0 for the population at large compared with SSBE only, but a broadly similar distribution curve. The UK-wide statistics stemming from EDA tally very well indeed with findings by Johns-Lewis [63], who reports a 128 Hz average for British English reading. With the acoustic phonetic corpus we provide a first account of how features such as speaking rate, voice quality, speech rhythm etc. are currently distributed in the UK speaker population which may be of service to forensic practitioners.

5. Challenges and benefits

There are some limitations and benefits to the EDAC, which we will discuss here. As shown in Section 3, the sampled corpora are skewed in so far as many of the macro-social categories do not reflect the distributions found in the UK population: there is an oversampling of a younger age group, male and females are not fully balanced, people with White ethnicity are overrepresented and people with higher education degrees are substantially oversampled. However, dialectology and variationist sociolinguistics have long tolerated certain biases in their methods of data collection. Not only traditional dialectological but also sociolinguistic methodologies have deeply entrenched notions of the types of 'authentic' speaker who should be the target of investigation: such 'authentic' speakers in variationist sociolinguistics are 'speakers of the pure vernacular' [64], i.e. a speaker who produces relaxed informal natural language in authentic, natural, contexts. Both, dialectology and variationist sociolinguistics, have been biased towards

population groups assumed to maintain the most distinctive regional or social varieties. Among traditional dialectologists, it was the Non-mobile Older Rural Males ('NORMs') ([1]; p. 30), among sociolinguists, usually non-mobile working class speakers of the 'vernacular' [64]. There are, however, benefits to the sampling applied in EDA protocol, which are discussed further below.

There are methodological issues which merit further discussion. The methods used in EDA are very different from those used in historical atlas work. For the SED, for example, researchers went into the field and conducted interviews. EDA was collected automatically and indirectly – with no researcher present. This leaves much less control over how the data were elicited; we do not know, for example, if users read the instructions properly. At the same time, this lack of human intervention in EDAC perhaps guards against the intrusion of experimenter bias and provides a degree of uniformity in the collection environment (cf. [25]). It is known, for example, that historical survey data collectors administered the task slightly differently, applying, for example, distinct transcription conventions [65].

Furthermore, EDA records read-aloud speech from the users – not spontaneously produced speech. Vernacular variants are bound to appear less in such a formal recording setting. The fact that we elicited read speech in the acoustic-phonetic corpus, too, somewhat limits the usefulness for forensic phonetics: in actual casework, most disputed samples (i.e. actual recordings from the perpetrators) are spontaneously produced. At the German Federal Criminal Police Office (BKA), for example, an estimated 80–90% of the disputed samples is produced spontaneously – not read off (Olaf Köster, BKA, personal communication). Spontaneous and read speech have been shown to differ on various linguistic levels; e.g. read speech has been shown to exhibit less

pauses, less disfluencies, and less reduction of vowels, but more f0 declination than spontaneous speech, to name just a few examples (cf. [66,67]). Some acoustic features are likely to be unaffected by speaking style, however, such as voice quality. These will be particularly useful for analyses and applicability for forensic phonetics in the future.

Another issue that deserves attention is the user's self-declared dialect and their selection of dialectal variants when going through the quiz. As described in section 2.1, users are asked to place the pin at the locality that best corresponds to their dialect. Here, we have to assume that users have an understanding both of their linguistic biography and of their language use (cf. [68]) – which we of course cannot investigate further with the data at hand. Further, when the users select their dialectal variants in the quiz (cf. 2.1) we ask them how they pronounce certain words. In doing so, we essentially collect people's intuitions about their language behavior, not people's actual language behavior. This is problematic – Labov [54] has shown, for example, that often people have erroneous intuitions when it comes to their actual language use, particularly with regard to socially salient variables. He observed, for example, that most people claim they say 'see you' at the end of a conversation; when in reality, it turned out that a majority said 'bye bye'. When confronted with this finding, speakers reacted in disbelief as 'bye bye' is frequently associated with childish behavior. It is further possible that users imitated a 'model' dialect when performing this task, perhaps because other variants are more prestigious. Or they may have nostalgically reported traditional variants that they no longer use.

Finally, it is worth pointing out that users are essentially performing a perception task (along with the help of IPA symbols), when asked to pick a variant from several variants that only differ in fine phonetic detail (e.g. 'room': [ɹʊm], [ɹʏm] or [ɹu:m]). At the moment we are testing the discrimination ability of naïve listeners, using the sound files embedded in the app: in a first pilot using an ABX paradigm, we tested 23 naïve listeners from England (12 SE, 6 NE, and 5 NW) and examined if they can discriminate items that included variants of th-fronting, glottalization, and velarization [69]. Results showed that, for glottalization, 73.5% could discriminate preglottalized, glottalized, and non-glottalized tokens, for th-fronting 91.2% (th-fronted, non-th-fronted), and for velarization 97.1% (velarized, vocalized, lateral). We concluded that (a) whether naïve listeners can perceive fine phonetic differences (as those included in the recordings in the App) depends on the token being examined and (b) the scores obtained are quite reassuring – even in tokens that are difficult to perceive (e.g. preglottalized, glottalized, non-glottalized tokens) we obtained high identification scores. A majority of naïve listeners are able to perceive the differences, thus validating the method. More generally, there are also technology-related limitations. We as researchers do not have contextual control over participants' physical environment – listeners may perform this perception task in a noisy environment. Vowel discrimination performance is likely to decrease with background noise [70] and discrimination performance is known to vary between listeners, depending on exposure to dialects, metalinguistic awareness, and location, amongst other factors [71,72]. Additionally, subjects may have submitted datasets multiple times. Perhaps, the same person used the app to participate repeatedly, or the same person used the app on a different smartphone, or 'performed' different dialects when responding to the questions in the quiz. To control for this, we included a button in the updated version that allows users to indicate if they had submitted the information before. Of those 11,181 users who evaluated the quiz result in the updated version, 6% had evaluated the result before – these sets were excluded from the statistics presented in section 3.1. For the 744 speakers who provided audio recordings in the updated version, 12% had provided recordings before – they were also excluded from the

descriptive statistics in section 3.2. For the large bulk of both corpora, thus, we cannot know for certain who submitted information multiple times. Reips [73], however, reports that the rate of repeated participants in internet-based testing is below 3% in most studies.

There are also obvious benefits to such a crowdsourced corpus. Concerning sampling, by radically changing data collection methods, EDAC gives up control over what constitutes the perfect 'authentic' speaker. Furthermore, among the rich metadata we are collecting from each user is information on personal mobility. This is of crucial value, as dialectology has traditionally attempted to exclude mobile individuals [65]. In this respect EDAC will enable a quantification of the effect of mobility on the phenomena under study – a blind spot of much previous work. Further, users appear to enjoy this type of playful exploration of dialects, which in return boosts the number of users that want to participate in the survey (cf. [74]). The corpus creation is comparatively cheap – conducting the current research using the same methods as applied in the SED would have been extremely costly.

Studies that use this type of automated data collection procedure require some form of quality control. One way of examining the validity is to compare trends found in EDAC with previously existing studies. In terms of the acoustic-phonetic corpus, we mentioned that the UK-wide f0 statistics found in EDAC male speakers align well with the values reported by Johns-Lewis [63] and with the distribution reported by [62]. As for the quiz corpus, the preliminary results we report on rhoticity (cf. Fig. 16), for example, reflect trends shown in previous studies which applied well-established dialectological methods, albeit being relatively localized and small scale (e.g. [75] for rhoticity in the South-West of England). We do not wish to claim that our approach should supplant existing techniques for the collection of dialectological data, of course, but simply wish to highlight the power, simplicity and added value of crowdsourced Big Data as a way of complementing established methods. As for data availability: the backend databases for the quiz and the acoustic-phonetic corpus have not yet been distributed to the general public. Once made public, the database can be used to aid future research in variationist sociolinguistics, forensic phonetics and phonetics more generally. The availability of EDAC to dialectologists and forensic phoneticians should lead to an improved understanding of the variability inherent in speech production and perception.

Acknowledgements

We thank Tam Blaxter from the University of Cambridge for his substantial contribution towards the English Dialects App project, Daniel Wanitsch of *ibros.ch* for technical app development, and Thomas Kettig for comments and edits that greatly improved the manuscript. We also show our gratitude to Sarah Grossenbacher and Melanie Calame for co-developing the app. Any errors are our own and should not tarnish the reputations of the persons acknowledged here. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.amper.2017.11.001>.

References

- [1] J. Chambers, P. Trudgill, *Dialectology*, second ed., Cambridge U. Press, Cambridge, 1998.
- [2] J.E. Schmidt, J. Herrgen, *Digitaler Wenker-atlas*, (2001) <http://www.diwa.info/>,

- Accessed date: 28 June 2017.
- [3] J. Gilliéron, *Atlas linguistique de la France*, Champion, Paris, 1902–1910.
 - [4] L. Anderwald, B. Szmezsanyi, *Corpus linguistics and dialectology*, in: A. Lüdeling, M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, 2009, pp. 1126–1139.
 - [5] A. Ellis, *On Early English Pronunciation: Part V*, Truebner and Co, London, 1889.
 - [6] L.L. Bonaparte, *On the dialects of Monmouthshire, Herefordshire, Worcestershire, Gloucestershire, Berkshire, Oxfordshire, South Warwickshire, South Northamptonshire, Buckinghamshire, Hertfordshire, Middlesex, and Surrey, with a new classification of the English dialects*, *Trans. Philol. Soc.* 16 (1) (1876) 570–579.
 - [7] W. Maguire, Mapping the existing phonology of English dialects, *Dialectol. Geolinguistica (DiG)* 20 (2012) 84–107.
 - [8] H. Kurath, G.S. Lowman, *The Dialectal Structure of Southern England: Phonological Evidence*, Publications of the American Dialect Society 54, U. Alabama Press, Tuscaloosa, 1970.
 - [9] H. Orton, E. Dieth, *Survey of English Dialects vol. 1*, E. J. Arnold & Son, Leeds, 1962.
 - [10] P. Trudgill, *The Social Differentiation of English in Norwich*, Cambridge University Press, Cambridge, 1974.
 - [11] D. Britain, When is a change not a change?: a case study on the dialect origins of New Zealand English, *Lang. Var. Change* 20 (2008) 187–223.
 - [12] H. Orton, S. Sanderson, J. Widdowson, *The Linguistic Atlas of England*, Croom Helm, London, 1978.
 - [13] R. Shackleton, *Quantitative Assessment of English-American Speech Relationships*, PhD Diss. Rijksuniversiteit Groningen, Groningen, 2010.
 - [14] L. Burnard, *Users Reference Guide British National Corpus Version 1.0*, Oxford U. Computing Services, Oxford, 1995.
 - [15] T. McNery, *Introducing a New Project with the British Library*, (2017) <http://cass.lancs.ac.uk/?cat=630/>, Accessed date: 28 June 2017.
 - [16] L. Anderwald, S. Wagner, FRED: the Freiburg English Dialect corpus, in: J. Beal, K.P. Corrigan, H.L. Moisl (Eds.), *Creating and Digitizing Language Corpora*, Vol. 1: *Synchronic Databases*, Palgrave Macmillan, London, 2007, pp. 35–53.
 - [17] K. Peitsara, A.-L. Vasko, *The Helsinki dialect corpus: characteristics of speech and aspects of variation*, *Hels. Engl. Studies* 2 (2002) Helsinki.
 - [18] L. MacKenzie, G. Bailey, D. Turton, *Crowdsourcing dialectology in the undergraduate classroom*, Paper presented at Methods in Dialectology XV, U. Groningen, 11–15.08.2014.
 - [19] B. Vaux, *Cambridge Online Survey of World Englishes*, (2017) http://www.tekstlab.uio.no/cambridge_survey/, Accessed date: 18 October 2017.
 - [20] D. Willis, *Using Twitter to investigate the diffusion of syntactic innovations*, Paper Presented at the Using Twitter for Linguistic Research Workshop, U. Kent, 31.05.2016.
 - [21] G. Bailey, *Regional variation in 140 characters: mapping geospatial tweets*, Paper Presented at the Using Twitter for Linguistic Research Workshop, U. Kent, 31.05.2016.
 - [22] J. Stevenson, *Mapping geographical variation in British English using Twitter*, Paper Presented at the Using Twitter for Linguistic Research Workshop, U. Kent, 31.05.2016.
 - [23] J. Grieve, C. Montgomery, A. Nini, *Assessing the use of social media for mapping lexical variation in British English*, *ICLAVE 9* (2017).
 - [24] M. Wieling, C. Upton, A. Thompson, *Analyzing the BBC Voices data: contemporary English dialect areas and their characteristic lexical variants*, *Lit. Ling. Comput.* 29 (1) (2013) 107–117.
 - [25] J. Godfrey, E.C. Holliman, J. McDaniel, *SWITCHBOARD: telephone speech corpus for research and development 1* (1992), pp. 517–520 ICASSP-92.
 - [26] A. Leemann, M.-J. Kolly, *Dialäkt App*, (2013) <https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8>, Accessed date: 28 May 2017.
 - [27] J. Katz, A. Wilson, *How y'all, youse and you guys talk*, *New York Times Online*, www.nytimes.com/interactive/2013/12/20/sunday-review/dialect-quiz-map.html?r=0, (20.12.2013), Accessed date: 28 June 2017.
 - [28] *New York Times web analytics group*, <http://www.nytc.co.com/wp-content/uploads/2013-Most-Visited-1.png>, (Accessed 28 June 2017).
 - [29] M.-J. Kolly, A. Leemann, *Dialäkt App: communicating dialectology to the public—crowdsourcing dialects from the public*, in: A. Leemann, M.-J. Kolly, V. Dellwo, S. Schmid (Eds.), *Trends in Phonetics in German-speaking Europe*, Peter Lang, Bern/Frankfurt, 2015, pp. 271–285.
 - [30] A. Leemann, M.-J. Kolly, J.-P. Goldman, V. Dellwo, I. Almajai, D. Wanitsch, *Voice App: a mobile app for crowdsourcing Swiss German dialect data*, *Proc. Interspeech 2015*, 2015, pp. 2804–2808.
 - [31] A. Leemann, M.-J. Kolly, Gruezi, Moin, Servus, (2015) <https://itunes.apple.com/ch/app/gruezi-moin-servus/id982633800?mt=8>, Accessed date: 28 June 2017.
 - [32] P. De Decker, J. Nycz, *For the record: which digital media can be used for socio-phonetic analysis? Working Papers in Ling.* vol. 17, U. Penn, 2011, pp. 51–59.
 - [33] C. Manfredi, J. Lebacqz, G. Cantarella, J. Schoentgen, S. Orlandi, A. Bandini, P.H. DeJonckere, *Smartphones offer new opportunities in clinical voice research*, *J. Voice* 31.1 (2017) 111–e1.
 - [34] *Stackoverflow.com, Sampling Rates of Smartphones*, <http://stackoverflow.com/questions/20889902/how-can-i-obtain-the-native-hardware-supported-audio-sampling-rates-in-order-t> (accessed 28.06.17).
 - [35] L.F. Lamel, R.H. Kassel, S. Seneff, *Speech database development: design and analysis of the acoustic-phonetic corpus*, *SIOA-1989*, 2 1989, pp. 161–170.
 - [36] D. Deterding, *The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation*, *JIPA*, vol. 36, 2006, pp. 187–196.
 - [37] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, *WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition*, *Proc. ICASSP 95*, 1995, pp. 81–84.
 - [38] ONS, *Population Estimates for UK, England and Wales, Scotland and Northern Ireland*, (2016) <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>, Accessed date: 28 June 2017.
 - [39] Statista.com, *Smartphone Ownership in the UK*, (2016) <http://www.statista.com/statistics/271851/smartphone-owners-in-the-united-kingdom-uk-by-age/>, Accessed date: 28 June 2017.
 - [40] ONS, *Ethnicity and National Identity in England and Wales: 2011*, (2016) <https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicityandnationalidentityinenglandandwales/2012-12-11>, Accessed date: 28 June 2017.
 - [41] ONS, *Highest Level of Qualification across England and Wales*, (2016) <http://www.ons.gov.uk/ons/about-ons/business-transparency/freedom-of-information/what-can-i-request/published-ad-hoc-data/census/qualifications/ct0429-2011-census-highest-level-of-qualification-england-and-wales.xls>, Accessed date: 28 June 2017.
 - [42] ONS, *What Qualification Levels Mean*, (2016) <https://www.gov.uk/what-different-qualification-levels-mean/overview>, Accessed date: 28 June 2017.
 - [43] ONS, *Distance Travelled to Work*, (2016) <http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/census/2011-census-analysis/distance-travelled-to-work/2011-census-analysis-distance-travelled-to-work.html>, Accessed date: 28 June 2017.
 - [44] T. Champion, *Urban-Rural differences in commuting in England: a challenge to the rural sustainability agenda?* *Plan. Pract. Res.* 24 (2009) 161–183.
 - [45] ONS, *Focus on People and Migration*, (2016) <http://www.ons.gov.uk/ons/rel/fertility-analysis/focus-on-people-and-migration/december-2005/focus-on-people-and-migration-focus-on-people-and-migration-chapters-6.pdf>, Accessed date: 28 June 2017.
 - [46] D. Van Leeuwen, F. Hinsken, B. Martinovic, A. van Hessen, S. Grondelaers, R. Orr, *Sprekend Nederland. A heterogeneous speech data collection*, *Comput. Ling. Neth. J.* 6 (2016) 21–38.
 - [47] University of Cambridge, *Cambridge app Maps Decline in Regional Diversity of English Dialects*, (2016) Press release <http://www.cam.ac.uk/research/news/cambridge-app-maps-decline-in-regional-diversity-of-english-dialects>, Accessed date: 28 June 2017.
 - [48] QGIS Development Team, *QGIS geographic information system. Open source geospatial foundation project*, <http://www.qgis.org/>, (2016), Accessed date: 28 June 2017.
 - [49] T. Blaxter, *Geospatial Temporal Visualisation*, U. Cambridge, Cambridge, 2017, pp. 101–120 PhD Diss. <https://doi.org/10.17863/CAM.15576>.
 - [50] S. Grossenbacher, *From East to West? Dialect Diffusion between Swindon and Bristol*, University of Bern, Bern, Switzerland, 2016 (MA Thesis).
 - [51] E. Valls, M. Wieling, J. Nerbonne, *Analyzing phonetic variation in the traditional English dialects: simultaneously clustering dialect and phonetic features*, *LLC J. Dig. Sch. Hum.* 28 (2013) 31–41.
 - [52] T. Kislir, U. Reichel, F. Schiel, C. Draxler, B. Jackl, N. Pörner, *BAS speech science web services: an update of current developments*, *Proc. LREC*, 10 2016, pp. 3880–3885.
 - [53] J. Harrington, S. Palethorpe, C.I. Watson, *Does the Queen speak the Queen's English?* *Nature* 408 (6815) (2000) 927–928.
 - [54] W. Labov, *When intuitions fail*, in: L. McNair, K. Singer, L.M. Dolbrin, M.M. Aucoin (Eds.), *Papers from the Parasession on Theory and Data in Linguistics*, Chicago Linguistic Society 32, Chicago, 1996, pp. 77–106.
 - [55] B. Vaux, S.A. Golder, *Harvard Dialect Survey*, (2003) <http://dialect.redlog.net/>, Accessed date: 28 June 2017.
 - [56] M. Calame, *Big Data in Variationist Sociolinguistics: From Tradition to Innovation*, University of Bern, Bern, Switzerland, 2016 (MA Thesis).
 - [57] E. Gold, P. French, *International practices in forensic speaker comparison*, *Int. J. Speech, Lang. Law* 18 (2) (2011) 293–307.
 - [58] O. Köster, R. Kehrein, K. Masthoff, Y.H. Boubaker, *The tell-tale accent: identification of regionally marked speech in German telephone conversations by forensic phoneticians*, *Int. J. Speech, Lang. Law* 19 (1) (2012) 51–71.
 - [59] M. Jessen, *Forensic phonetics*, *Lang. Ling. Compass* 2 (2008) 671–711.
 - [60] P. French, P. Harrison, *Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases*, *Intl. J. Speech, Lang. Law* 14 (2007) 137–144.
 - [61] T. Hudson, A. Leemann, M.-J. Kolly, D. Britain, K. McDougall, *Preliminary crowdsourced UK fundamental frequency population data for English speakers*, Paper presented at IAFPA 2016, U. York, York, 24–27.07.2016.
 - [62] T. Hudson, G. de Jong, K. McDougall, P. Harrison, F.J. Nolan, *F0 statistics for 100 young male speakers of Standard Southern British English*, *Proc. ICPhS*, 16 2007, pp. 1809–1812.
 - [63] C. Johns-Lewis, *Prosodic differentiation of discourse modes*, in: C. Johns-Lewis (Ed.), *Intonation in Discourse*, Croom Helm, London, 1986, pp. 199–219.
 - [64] P. Eckert, *Elephants in the room*, *J. Socioling.* 7 (2003) 392–397.

- [65] D. Britain, Between North and South: the Fenland, in: R. Hickey (Ed.), *Researching Northern English*, Benjamins, Amsterdam, 2015, pp. 417–436.
- [66] G.P.M. Laan, The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style, *Speech Comm.* 22 (1997) 43–65.
- [67] E. Shriberg, Spontaneous speech: how people really talk, and why engineers should care, *Proc. Eur. Conf. Speech Comm. Tech.* 9 2005, pp. 1781–1784.
- [68] A. Leemann, M.-J. Kolly, R. Purves, D. Britain, E. Glaser, Crowdsourcing language change with smartphone applications, *PLoS One* 11 (1) (2016).
- [69] A. Leemann, Using smartphone apps to map phonetic variation in British English, German, and Swiss German, *J. Acoust. Soc. Am.* (2017) 141 3909.
- [70] G. Parikh, P.C. Loizou, The influence of noise on vowel and consonant cues, *J. Acoust. Soc. Am.* 118 (6) (2005) 3874–3888.
- [71] M. Cooke, J. Barker, M.L.G. Lecumberri, Crowdsourcing in speech perception, in: M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, D. Suendermann (Eds.), *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, John Wiley & Sons, Hoboken, 2013, pp. 137–172.
- [72] J. Hay, R. Podlubny, K. Drager, M. McAuliffe, Car-talk: location-specific speech production and perception, *J. Phon.* 65 (2017) 94–109.
- [73] U.-D. Reips, Standards for internet-based experimenting, *Exp. Psychol.* 49 (2002) 243–256.
- [74] G. Parent, M. Eskenazi, Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges, *Proc. Interspeech* 2011, 2011, pp. 3037–3040.
- [75] C. Piercy, A transatlantic, cross-dialectal comparison of non-pre-vocalic/r/, *Working Papers in Ling.* 18 U. Penn, 2012, pp. 77–86.